Hodnocení diplomové práce

Jméno studenta:	**Bc. Tomáš Lindauer**
Téma diplomové práce:	**Statistická analýza výsledků sportovních soutěží**
Hodnotitel – oponent práce:	Ing. Patrice Marek, Ph.D. Západočeská univerzita v Plzni Fakulta aplikovaných věd

---

Předložená práce se zabývá analýzou výsledků v ledním hokeji a zaměřuje se na dvě hlavní témata – rozdílnost mezi soutěžemi a analýzu průběhu zápasů.

Na úvodních stranách se autor věnuje popisu dat a změn bodovacích systémů v průběhu let. Následně je zkoumána výhodnost domácího prostředí. Ve třetí kapitole se autor věnuje averzi ke ztrátě, kde prezentuje teorii a formuluje v návaznosti na teorii hypotézy spojené s hokejovým utkáním, které jsou testovány ve čtvrté kapitole práce.

V páté kapitole je otestován vliv systému bodování a šestá kapitola se zabývá dynamikou zápasu ve třetí třetině zápasu.

Hodnocení obsahové stránky

Předložená práce obsahuje mnoho chyb, často velmi zásadních. V celé práci jde především o testování hypotéz, kdy k většině z nich lze mít výhrady. Zpracování testů a modelů působí tak, že autor na data „pustil" program a vypsal závěr programu. Přitom se nevěnoval splnění předpokladů, vhodnosti použitých testů a modelů a v mnoha případech ani přesné definici nulové a alternativní hypotézy.

V následujícím textu uvádím přehled chyb. Zásadní chyby jsou odlišeny tučným textem.

- V prezentovaných statistikách (např. str. 8) je použit medián pro diskrétní data – autor se zde vůbec nevěnuje vysvětlení problematiky stanovení mediánu pro diskrétní data.
- Na straně 10 je uvedeno, že v práci nebudou uvažovány poslední 3 minuty zápasů, kvůli nestandardní herní situaci. Bylo by vhodné podložit toto tvrzení i testy. O poslední minutě to lze tvrdit i bez testů, ovšem 58. a 59. minuta již tak jasně nevypadají. V páté kapitole je pak uvažován průběh s vyloučením posledních 5 minut zápasu. Tato změna není nijak odůvodněna. Uvedené grafy (např. str. 11) naznačují, že problematická bude i první minuta, tím se autor ale nezabýval.
- Na straně 13 je stanovena nulová hypotéza takto: „pravděpodobnost výhry v základní hrací době pro domácí tým je v NHL stejná jako ve všech ostatních soutěžích". **Hlavním problémem je, že pro testování hypotézy je nevhodně zvolen test nezávislosti v kontingenční tabulce. Vzhledem k charakteru Chí-kvadrát rozdělení je nesprávné i tvrzení, že p-hodnota testu je rovna nule.** Nelze také tvrdit, že jde o „všechny ostatní soutěže", ale o „všechny *ostatní testované* soutěže". K použitému testu není prezento-

vána žádná teorie, pouze je uveden odkaz na literaturu, kde je možné tento test nalézt. Toto nepovažuji za vhodné řešení, naprosté základy testu by vždy měly být uváděny.

- Obrázek na straně 16 rozhodně nezobrazuje pravděpodobnostní funkci. Stejně tak i obrázek na straně 17 by neměl obsahovat spojité průběhy.
- Ve třetí kapitole jsou k prezentaci použity převzaté obrázky, což není zdůrazněno. Obrázky jsou převzaty i s původním anglickým titulkem a anglickými popisky (str. 19 a 21). Bylo by vhodnější vytvořit tyto jednoduché obrázky samostatně.
- **V části 4.1 je stanovena nulová hypotéza, že v každé třetině padá stejný počet gólů. Test je proveden naprosto nesmyslně. Jsou zde sice správně stanoveny očekávané hodnoty počtu gólů, ale následně je nad touto tabulkou pozorovaných a očekávaných hodnot proveden chí-kvadrát test nezávislosti, místo toho aby byl použit chí-kvadrát test dobré shody. Toto se opakuje i u následujících dvou testů (str. 26 a 27).**
- **V další podkapitole se autor věnuje regresi, pomocí které se snaží otestovat hypotézu o nezávislosti počtu branek na minutě zápasu. Jde opět o nevhodný postup, zde by bylo na místě testovat shodu s rovnoměrným rozdělením, tj. že v každé minutě padá stejný počet branek. Místo toho je daty proložena přímka a je testována nulová hypotéza, zda je směrnice přímky rovna nule. V případě, že se nulovou hypotézu povede zamítnout, tak je přijata alternativa, že průměrný počet branek je závislý na minutě zápasu. Zde stačí jako protipříklad uvést situaci, kdy v každé liché minutě padá v průměru jeden gól a v každé sudé minutě nepadne žádný gól. Je zde zcela jasné, že zde závislost existuje, ale při použití lineární regrese nebude nulová hypotéza zamítnuta (tj. nepovede se zamítnout hypotézu o nulovosti směrnice).**

  Krom uvedeného se autor nevěnuje uvedení teorie spojené s lineární regresí, a to ani odkazem na použitou literaturu. Chybí jakékoliv ověření předpokladů pro použití lineární regrese i komentář o tom, jak byl odhad proveden. Z obrázku na straně 25 se lze domnívat, že data obsahují odlehlá pozorování a použití lineární regrese bude tedy problematické. Samozřejmě, otevřenou otázkou je i splnění předpokladů na náhodnou složku.

  V návaznosti na předchozí text je zajímavé i vyvození následujícího závěru na straně 28: „za zmínku stojí výsledek lineární regrese pro druhou třetinu švédské ligy, kde jako v jediném případě se s přibývajícím časem průměrný počet branek snižuje. Tato závislost se ale ukázala jako statisticky nevýznamná". Obě tvrzení se dle mého názoru vylučují.
- **V následující kapitole je použita nelineární regrese k proložení dat. Chybí zde uvedení jakékoliv motivace k použití tohoto modelu**. Z mého pohledu není model vhodný, jelikož nevystihuje charakter dat (zde se ale jedná pouze o můj subjektivní pohled). **Problémem je opět chybějící teorie (opět ani odkazem), dále pak chybějící chybový člen v modelu, chybějící požadavky na model i použitá metoda odhadu (konstatování, že odhad byl proveden v programu Matlab považuji za nedostatečné).**

- Jak již bylo zmíněno, od páté kapitoly je bez zdůvodnění brán v úvahu čas pouze do 55. minuty, přičemž předchozí části práce uvažují čas až do 57. minuty. Při analýze vlivu změny systému bodování není ani zmínka o možnosti vlivu jiných faktorů, např. posunutí modré čáry, změna pravidel pro velikosti výstroje brankáře apod.
- V tabulkách od strany 34 je uváděna „směrodatná odchylka", mělo by ale být uvedeno „výběrová směrodatná odchylka".
- **Pro testování shody středních hodnot dvou výběrů je použit t-test pro dva nezávislé výběry. Opět není uvedena žádná teorie, pouze odkaz na literaturu a schází jakékoliv vysvětlení použití tohoto testu, kdy první volbou by měl být t-test pro dva nezávislé výběry z normálních rozdělení se stejnými rozptyly. Teprve po nesplnění předpokladů by měl být použit obecnější test. Test na straně 34 je test uveden jako oboustranný, ale je prováděn jako jednostranný (viz p-hodnotu testu pro daná data).**
- U provedených testů na straně 40 a 42 není žádná poznámka o odlišnosti stanovení p-hodnoty pro výběry o malém rozsahu.
- Na straně 41 je testován vliv změny bodování na „neremízová" utkání po druhé třetině, tj. kdy jeden z týmů vede alespoň o jeden gól. Zápasy jsou zde rozděleny do skupin o 2, 4, 6, 8 a 10 a více gólech. Jsou zde tedy bez udání důvodu ignorovány stavy zápasů s lichým počtem gólů, kdy samozřejmě jeden z týmů vede.
- Na straně 43 je stanovena nulová hypotéza o závislosti takto: „padnutí branky ve 3. třetině nezávisí na bodovém systému". Stanovení směru závislosti je vždy problematické, proto by mělo být správně uvedeno, že je testována závislost mezi veličinami. Navíc, ve formulaci hypotézy by se měla objevit podmínka o stavu 1:1, který je předpokládán po druhé třetině.
- V šesté části se autor zabývá analýzou rozptylu. Jsou zde oproti předchozím částem otestovány předpoklady. K otestování homoskedasticity je zvolen Leveneův test. Znovu ale platí, že zde test není popsán ani nijak komentován, není ani stanoven důvod výběru tohoto testu.
- **Pro test normality na straně 47 je zvolen Lillieforsův test, opět bez udání důvodu pro výběr tohoto testu, uvedení teorie nebo alespoň odkazu na literaturu. Autor sice zamítne nulovou hypotézu o tom, že data jsou normálně rozdělena, ale odvolává se na centrální limitní větu a tvrdí: „pokud jsou ale soubory dostatečně rozsáhlé ($n_i > 50$), tak se data na základě centrální limitní věty chovají přibližně jako z normálního rozdělení". Toto je nesmysl, jelikož centrální limitní věta se týká průměru a součtu z náhodných veličin, navíc vyžaduje dodatečné předpoklady.**
- **ANOVA není vysvětlena, také není zdůvodněno, proč byl tento model zvolen, proč jsou zanedbány interakce apod. Jak bylo uvedeno dříve, data nejsou normálně rozdělena (jsou dokonce diskrétní – jde o počet gólů), tak jak to je pro tento test vyžadováno. Pokládám použití tohoto testu za nevhodné.**

- **Navržená modifikace na straně 54 obsahuje průměrný počet gólů, ale jak bylo uvedeno, autor pracuje s počtem gólů, není tedy jasné, jakým způsobem získá pozorování průměrných počtů.**

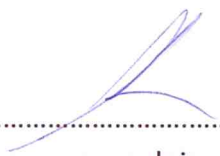
Hodnocení formálního zpracování

Grafické zpracování je na první pohled pečlivé a téměř bez překlepů či jiných chyb. Problémy jsou již uvedeny v předchozím textu, tj. zobrazení pravděpodobnostní funkce na str. 16, dále pak přejaté obrázky (str. 19, 21 a 22) a vynechané místo (např. str. 12, str. 20).

Na nízké úrovni je způsob prezentace. Jedná se o práci z oblasti statistiky, kde je nutné dbát na správné formulace. Především to platí o formulaci hypotéz a vyvozování správných závěrů. V práci by se také nemělo objevit prohlášení, že je přijímána nulová hypotéza (str. 40 uprostřed a 46 uprostřed).

Vzhledem k uvedeným skutečnostem ***nedoporučuji*** předloženou diplomovou práci k obhajobě před státní komisí a navrhuji tedy hodnocení ***nevyhověl***.

V Plzni dne 5. 6. 2012



.....

podpis